

An Inequality with Applications to Structured Sparsity and Multitask Dictionary Learning

Andreas Maurer

AM@ANDREAS-MAURER.EU

Adalbertstrasse 55, D-80799 Munchen, Germany

Massimiliano Pontil

M.PONTIL@CS.UCL.AC.UK

Department of Computer Science

Centre for Computational Statistics and Machine Learning

University College London, UK

Bernardino Romera-Paredes

BERNARDINO.PAREDES.09@UCL.AC.UK

Department of Computer Science and UCL Interactive Centre

University College London, UK

Abstract

From concentration inequalities for the suprema of Gaussian or Rademacher processes an inequality is derived. It is applied to sharpen existing and to derive novel bounds on the empirical Rademacher complexities of unit balls in various norms appearing in the context of structured sparsity and multitask dictionary learning or matrix factorization. A key role is played by the largest eigenvalue of the data covariance matrix.

Keywords: Concentration inequalities, multitask learning, Rademacher complexity, risk bounds, structured sparsity.

1. Introduction

The method of Rademacher complexities (Bartlett and Mendelson, 2002; Koltchinskii and Panchenko, 2002) has become a standard tool to prove generalization guarantees for learning algorithms. One considers a loss class \mathcal{F} of functions $f : \mathcal{X} \rightarrow \mathbb{R}$, where \mathcal{X} is some space of examples (such as input-output pairs), a sample $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$ of observations and a vector $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ of independent Rademacher variables ϵ_i uniformly distributed on $\{-1, 1\}$. The Rademacher complexity $\mathcal{R}(\mathcal{F}, \mathbf{x})$ is then defined as

$$\mathcal{R}(\mathcal{F}, \mathbf{x}) = \frac{2}{n} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i f(x_i). \quad (1)$$

Bounds on Rademacher complexities are useful in learning theory because they lead to uniform bounds, as for example in the following result (Bartlett and Mendelson, 2002).

Theorem 1 *Suppose the members of \mathcal{F} take values in $[0, 1]$, let X, X_1, \dots, X_n be iid random variables with values in \mathcal{X} , and let $\mathbf{X} = (X_1, \dots, X_n)$. Then for $\delta > 0$ with probability at least $1 - \delta$ we have for every $f \in \mathcal{F}$ that*

$$\mathbb{E} f(X) \leq \frac{1}{n} \sum_{i=1}^n f(X_i) + \mathcal{R}(\mathcal{F}, \mathbf{X}) + \sqrt{\frac{9 \ln 2/\delta}{2n}}.$$

Since also for any real L -Lipschitz function ϕ we have $\mathcal{R}(\phi \circ \mathcal{F}, \mathbf{x}) \leq L \mathcal{R}(\mathcal{F}, \mathbf{x})$ (see e.g. [Bartlett and Mendelson, 2002](#)) the utility of Rademacher complexities is not limited to functions with values in $[0, 1]$.

For many function classes \mathcal{F} considered in machine learning one can find other function classes $\mathcal{F}_1, \dots, \mathcal{F}_M$ such that

$$\sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i f(x_i) \leq \max_{m=1}^M \sup_{f \in \mathcal{F}_m} \sum_{i=1}^n \epsilon_i f(x_i). \quad (2)$$

Multiple kernel learning (see e.g. [Bach et al., 2005](#); [Cortes et al., 2010](#); [Ying and Campbell, 2009](#)) provides an example. Let $\mathcal{H} = \mathcal{H}_1 \oplus \dots \oplus \mathcal{H}_M$ be the direct sum of Hilbert spaces \mathcal{H}_m with norm $\|\cdot\|$ and inner product $\langle \cdot, \cdot \rangle$. The \mathcal{H}_m are reproducing kernel Hilbert spaces induced by kernels κ_m with corresponding feature maps $\psi_m : \mathcal{X} \rightarrow \mathcal{H}_m$. We denote by $\psi = (\psi_1, \dots, \psi_M) : \mathcal{X} \rightarrow \mathcal{H}$ the composite feature map and define the group norm for $\beta = (\beta_1, \dots, \beta_M) \in \mathcal{H}$ by

$$\|\beta\|_G = \sum_{m=1}^M \|\beta_m\|.$$

We are interested in the class of functions $\mathcal{F} = \{x \in \mathcal{X} \mapsto \langle \psi(x), \beta \rangle : \|\beta\|_G \leq 1\}$. It is easy to see that the dual norm to $\|\cdot\|_G$ is $\|z\|_{G,*} = \max_m \|z_m\|$. We therefore have, writing $\mathcal{F}_m = \{x \mapsto \langle \psi_m(x), \beta \rangle_m : \beta \in \mathcal{H}_m, \|\beta\| \leq 1\}$,

$$\sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i f(x_i) = \left\| \sum_{i=1}^n \epsilon_i \psi(x_i) \right\|_{G,*} = \max_{m=1}^M \sup_{f \in \mathcal{F}_m} \sum_{i=1}^n \epsilon_i f(x_i),$$

as in (2), so $\mathcal{R}(\mathcal{F}, \mathbf{x}) \leq \mathcal{R}(\cup_m \mathcal{F}_m, \mathbf{x})$. In the sequel we show that many classes encountered in the study of structured sparsity, matrix factorization and multitask dictionary learning allow similar decompositions.

This paper proposes a simple general method to obtain uniform bounds for these cases and applies it to sharpen some existing ones, and to derive some new results. The method is based on the following innocuous looking lemma.

Lemma 2 *Let $M \geq 4$ and $A_1, \dots, A_M \subset \mathbb{R}^n$, $A = \cup_m A_m$, and let $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ be a vector of independent Rademacher variables. Then*

$$\mathbb{E} \sup_{\mathbf{z} \in A} \langle \epsilon, \mathbf{z} \rangle \leq \max_{m=1}^M \mathbb{E} \sup_{\mathbf{z} \in A_m} \langle \epsilon, \mathbf{z} \rangle + 4 \sup_{\mathbf{z} \in A} \|\mathbf{z}\| \sqrt{\ln M}.$$

If the ϵ_i are replaced by standard normal variables the same conclusion holds with the constant 4 replaced by 2.

For function classes $\mathcal{F}_1, \dots, \mathcal{F}_M$ and a sample \mathbf{x} let A_m be the subset of \mathbb{R}^n defined by $A_m = \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}_m\}$ to see that the conclusion reads

Corollary 3 *Let $\mathfrak{S} = \max_{m=1}^M \mathcal{R}(\mathcal{F}_m, \mathbf{x})$ and let $\mathfrak{W} = \sqrt{\sup_{f \in \cup \mathcal{F}_m} \frac{1}{n} \sum_{i=1}^n f^2(x_i)}$. Then*

$$\mathcal{R}\left(\bigcup_{m=1}^M \mathcal{F}_m, \mathbf{x}\right) \leq \mathfrak{S} + 8\mathfrak{W} \sqrt{\frac{\ln M}{n}}.$$

To apply this inequality we have to bound the strong parameter \mathfrak{S} and the weak parameter \mathfrak{W} . The real advantage of the trick lies in the weak parameter which becomes small if the function classes have a high linguistic specificity in the sense that individual functions are appreciably different from zero only for rather special types of inputs. In the context of linear prediction this corresponds to a small spectral norm of the covariance operator, a phenomenon often associated with high dimensionality (see Section 2.2.). In such cases the complexity of the most complex class becomes the dominant term in the bound.

For multiple kernel learning we find with standard methods

$$\mathfrak{S} \leq \frac{2}{n} \max_{m=1}^M \sqrt{\sum_{i=1}^n \|\psi_m(x)\|^2} = 2 \max_{m=1}^M \sqrt{\frac{\text{tr}(\hat{C}(\psi_m(\mathbf{x})))}{n}}$$

where the uncentered empirical covariance operator of the data $\mathbf{z} = (z_1, \dots, z_n)$ is defined, for every vectors v, w , by the equation $\langle \hat{C}(\mathbf{z})v, w \rangle = \frac{1}{n} \sum_{i=1}^n \langle v, z_i \rangle \langle z_i, w \rangle$, see also Section 2.2 below. The weak parameter is

$$\mathfrak{W} = \max_{m=1}^M \sqrt{\sup_{\beta \in \mathcal{F}_m} \frac{1}{n} \sum_{i=1}^n \langle \beta, \psi_m(x_i) \rangle^2} = \max_{m=1}^M \sqrt{\lambda_{\max}(\hat{C}(\psi_m(\mathbf{x})))},$$

where λ_{\max} denotes the largest eigenvalue. The overall bound is then

$$\mathcal{R}(\mathcal{F}, \mathbf{x}) \leq 2 \max_{m=1}^M \sqrt{\frac{\text{tr}(\hat{C}(\psi_m(\mathbf{x})))}{n}} + 8 \max_{m=1}^M \sqrt{\frac{\lambda_{\max}(\hat{C}(\psi_m(\mathbf{x}))) \ln M}{n}}. \quad (3)$$

Note that in this example the eigenvalues of $\hat{C}(\psi_m(\mathbf{x}))$ coincide with the eigenvalues of the normalized kernel matrix $\kappa_m(x_i, x_j)/n$. Other authors ([Cortes et al., 2010](#); [Maurer and Pontil, 2012](#)) give a bound of order $\max_{m=1}^M \sqrt{\text{tr}(\hat{C}(\psi_m(\mathbf{x}))) \ln M/n}$. If we divide the two bounds we see that (3) becomes a significant improvement when the number of kernels is large and the quotient $\lambda_{\max}(\hat{C}(\psi_m(\mathbf{x})))/\text{tr}(\hat{C}(\psi_m(\mathbf{x})))$ is small. The latter condition will occur if the feature representations $\psi_m(\mathbf{x})$ are essentially high dimensional, as it occurs for example with Gaussian radial basis function kernels with small kernel width. This type of behaviour is typical of the proposed method whose benefits become more pronounced in effectively high dimensions. In the artificial case of exactly spherical data \mathbf{x} in \mathbb{R}^d , we even have $\lambda_{\max}(\hat{C}(\mathbf{x}))/\text{tr}(\hat{C}(\mathbf{x})) = 1/d$ (see Section 2.2).

Of course the example of multiple kernel learning applies equally to the group lasso ([Yuan and Lin, 2006](#)), but Lemma 2 can also be applied to a large class of structured sparsity norms to sharpen bounds for overlapping groups ([Jacob et al., 2009](#)), cone regularizers ([Micchelli et al., 2013](#)) and the recently proposed k -support norm ([Argyriou et al., 2012](#)).

Related applications give generalization guarantees for various schemes of multitask dictionary learning or matrix factorization. As examples we reproduce the results by [Maurer et al. \(2013\)](#) and give novel bounds for other matrix regularizers including multitask subspace learning. In these applications the weak parameter is particularly important, because it is proportional to the limit of the generalization error as the number of tasks goes to infinity.

The proof of Lemma 2 relies on concentration inequalities for the suprema of Rademacher or Gaussian processes. If the random variables ϵ_i are independent standard normal then the constant 4 in Lemma 2 can be replaced by 2. On the other hand bounding the Rademacher complexities by Gaussian complexities incurs a factor of $\sqrt{\pi/2}$, so little seems to be gained. We will however also give the bound for isonormal ϵ because Gaussian averages are sometimes convenient when Slepian's inequality is applied to simplify complicated classes.

Lemma 2 is certainly not new, although we cannot give an exact reference. Related results appear in various disguises whenever modern concentration inequalities are applied to empirical processes, as for example in (Ledoux and Talagrand, 1991) or the recent book by Boucheron et al. (2013). We are not aware of any reference where Lemma 2 is applied as a systematic method to prove or improve uniform bounds as in the present paper. The applications given are intended as illustrations of the method and they are by no means exhaustive. The bound in Section 3.3 has already appeared in (Maurer et al., 2013), the result on subspace learning in Section 3.5 is somewhat similar to a result derived from noncommutative Bernstein inequalities in (Maurer and Pontil, 2013). The bounds on structured sparsity norms in Section 3.1 and the result for the sharing norm in Section 3.4 are new to the best of our knowledge.

In the next section we give a proof of Lemma 2 and in Section 3 we give applications to structured sparsity and dictionary learning. An appendix contains the proofs of the concentration inequalities we use.

2. Theory

We provide a proof of Lemma 2 and a brief and elementary discussion of covariances.

2.1. The Proof of Lemma 2

We use the following concentration inequality for the suprema of bounded or Gaussian random processes. A proof and bibliographical remarks are provided in the technical appendix to this paper.

Theorem 4 *Let $A \subset \mathbb{R}^n$ and let $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ be a vector of independent random variables satisfying $|\epsilon_i| \leq 1$. Then*

$$\Pr \left\{ \sup_{\mathbf{z} \in A} \langle \epsilon, \mathbf{z} \rangle > \mathbb{E} \sup_{\mathbf{z} \in A} \langle \epsilon, \mathbf{z} \rangle + s \right\} \leq \exp \left(\frac{-s^2}{8 \sup_{\mathbf{z} \in A} \|\mathbf{z}\|^2} \right).$$

If the ϵ_i are replaced by standard normal variables then the same conclusion holds and the constant 8 can be replaced by 2.

With Theorem 4 at hand the proof of Lemma 2 becomes an exercise of calculus.

Proof of Lemma 2. Denote the random variable $\sup_{\mathbf{z} \in A_m} \langle \epsilon, \mathbf{z} \rangle$ with F_m and set $v = \sup_{\mathbf{z} \in A} \|\mathbf{z}\|$. From Theorem 4 we have for $s > 0$

$$\Pr \{F_m > \mathbb{E}F_m + s\} \leq e^{-s^2/(2b^2v^2)},$$

where b is either 1 in the Gaussian or 2 in the bounded case. A union bound gives

$$\Pr \left\{ \max_m F_m > \max_m \mathbb{E} F_m + s \right\} \leq M e^{-s^2/(2b^2v^2)}. \quad (4)$$

We now have, for any positive δ ,

$$\begin{aligned} \mathbb{E} \max_m F_m &\leq \max_m \mathbb{E} F_m + \delta + \int_{\max_m \mathbb{E} F_m + \delta}^{\infty} \Pr \left\{ \max_m F_m > s \right\} ds \\ &= \max_m \mathbb{E} F_m + \delta + \int_{\delta}^{\infty} \Pr \left\{ \max_m F_m > \max_m \mathbb{E} F_m + s \right\} ds \\ &\leq \sup_m \mathbb{E} F_m + \delta + M \int_{\delta}^{\infty} e^{-s^2/(2b^2v^2)} ds. \end{aligned}$$

The first step holds because probabilities do not exceed one, the second is a change of variable and finally we used (4). By a well known approximation we can bound the integral by

$$\int_{\delta}^{\infty} e^{-s^2/(2b^2v^2)} ds \leq \frac{b^2v^2}{\delta} e^{-\delta^2/(2b^2v^2)}.$$

Using $\delta = \sqrt{2b^2v^2 \ln M}$ we have

$$\begin{aligned} \mathbb{E} \max_m F_m &\leq \max_m \mathbb{E} F_m + \delta + \frac{Mb^2v^2}{\delta} e^{-\delta^2/(2b^2v^2)} \\ &= \max_m \mathbb{E} F_m + bv \left(\sqrt{2 \ln M} + \frac{1}{\sqrt{2 \ln M}} \right) \leq \max_m \mathbb{E} F_m + 2bv\sqrt{\ln M}, \end{aligned}$$

since we assumed $M \geq 4$. ■

2.2. Covariances

Let $\mathbf{x} = (x_1, \dots, x_n)$ be a sequence of points in a finite or infinite dimensional real Hilbert space \mathcal{H} with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. The (uncentered, empirical) covariance operator $\hat{C}(\mathbf{x})$ is defined by

$$\left\langle \hat{C}(\mathbf{x}) v, w \right\rangle = \frac{1}{n} \sum_{i=1}^n \langle x_i, v \rangle \langle x_i, w \rangle, \quad v, w \in \mathcal{H}.$$

$\hat{C}(\mathbf{x})$ is positive semidefinite and of rank at most n . Its trace is given by

$$\text{tr} \left(\hat{C}(\mathbf{x}) \right) = \frac{1}{n} \sum_{i=1}^n \|x_i\|^2.$$

In the sequel we will frequently use the inequality

$$\mathbb{E} \left\| \sum_{i=1}^n \epsilon_i x_i \right\| \leq \left(\mathbb{E} \left\| \sum_{i=1}^n \epsilon_i x_i \right\|^2 \right)^{1/2} = \sqrt{n \text{tr} \left(\hat{C}(\mathbf{x}) \right)} \quad (5)$$

where the ϵ_i are either independent Rademacher or standard normal variables and we used Jensen's inequality and the orthonormality properties of the ϵ_i .

The largest eigenvalue of the covariance is

$$\lambda_{\max}(\hat{C}(\mathbf{x})) = \sup_{\|v\| \leq 1} \langle \hat{C}(\mathbf{x}) v, v \rangle = \sup_{\|v\| \leq 1} \frac{1}{n} \sum_i \langle x_i, v \rangle^2.$$

Clearly the ratio $\lambda_{\max}(\hat{C}(\mathbf{x}))/\text{tr}(\hat{C}(\mathbf{x}))$ is upper bounded by 1 and it can be as small as $1/n$ for exactly spherical data.

For a practical example suppose that the inputs lie in \mathbb{R}^d and that the Hilbert space \mathcal{H} is induced by a Gaussian kernel, so that

$$\langle \psi(x), \psi(y) \rangle = \kappa(x, y) = \exp\left(-\frac{\|x - y\|_{\mathbb{R}^d}^2}{\sigma^2}\right),$$

where ψ is the embedding feature map and $\|\cdot\|_{\mathbb{R}^d}$ is the standard inner product in \mathbb{R}^d . Now let a dataset $\mathbf{x} = (x_1, \dots, x_n)$ be given with $x_i \in \mathbb{R}^d$. Clearly $\text{tr}(\hat{C}(\psi(\mathbf{x}))) = 1$. Suppose that Δ is the smallest distance between any two observations $\Delta = \min_{i \neq j} \|x_i - x_j\|_{\mathbb{R}^d}$. It is easy to see that the largest eigenvalue of the covariance is $1/n$ times the largest eigenvalue of the kernel matrix $K = \kappa(x_i, x_j)_{i,j=1}^n$. Thus

$$\begin{aligned} \lambda_{\max}(\hat{C}(\psi(\mathbf{x}))) &= \frac{1}{n} \sup_{\|\alpha\|_{\mathbb{R}^d} \leq 1} \langle K\alpha, \alpha \rangle = \frac{1}{n} + \frac{1}{n} \sup_{\|\alpha\|_{\mathbb{R}^d} \leq 1} \sum_{i \neq j} \alpha_i \alpha_j \exp\left(-\frac{\|x_i - x_j\|_{\mathbb{R}^d}^2}{\sigma^2}\right) \\ &\leq \frac{1}{n} + \frac{e^{-\Delta^2/\sigma^2}}{n} \sup_{\|\alpha\|_{\mathbb{R}^d} \leq 1} \sum_{i \neq j} |\alpha_i| |\alpha_j| \leq \frac{1}{n} + e^{-\Delta^2/\sigma^2}. \end{aligned}$$

This is also a bound on the ratio $\lambda_{\max}(\hat{C}(\mathbf{x}))/\text{tr}(\hat{C}(\mathbf{x}))$, since the trace of the covariance is 1 for the Gaussian kernel. It follows that the weak parameter in our bounds decreases with the width σ of the kernel. Of course this is only part of the story. We hasten to add that decreasing the kernel width will have an adverse effect on generalization. Nevertheless our results seem to indicate that, at least in the context of the applications below, the kernel width can be chosen smaller than suggested by conventional bounds, where λ_{\max} is replaced by the trace (Maurer and Pontil, 2012; Kakade et al., 2012; Cortes et al., 2010). This is particularly true for multitask learning with a large number of tasks, where λ_{\max} scales the limiting generalization error, as shown below.

We state our bounds in terms of uncentered covariances, but of course they also apply as well if the data is centered by subtracting $\bar{\mathbf{x}} = (1/n) \sum_i x_i$ from each data point. It is easy to see that $\langle \hat{C}(\mathbf{x} - \bar{\mathbf{x}}) v, v \rangle \leq \langle \hat{C}(\mathbf{x}) v, v \rangle$ for all v , so that $\text{tr}(\hat{C}(\mathbf{x} - \bar{\mathbf{x}})) \leq \text{tr}(\hat{C}(\mathbf{x}))$ and $\lambda_{\max}(\hat{C}(\mathbf{x} - \bar{\mathbf{x}})) \leq \lambda_{\max}(\hat{C}(\mathbf{x}))$. Our bounds can therefore only benefit from centering. This is relevant when calculating the advantage of our bounds in practice. With MNIST and raw pixel data without kernel, we found $\lambda_{\max}(\hat{C}(\mathbf{x}))/\text{tr}(\hat{C}(\mathbf{x})) \approx 0.95$ for uncentered data, but < 0.1 for centered data.

3. Application Examples

We use Lemma 2 to derive general bounds for a class of structured sparsity norms. Then we discuss several applications to multitask dictionary learning.

3.1. Structured Sparsity

Suppose \mathcal{H} is a separable, real, finite or infinite dimensional Hilbert space with norm and inner product $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$, and that $\mathcal{P} = \{P_1, \dots, P_M\}$ is a collection of symmetric bounded operators whose ranges together span \mathcal{H} . We consider the infimal convolution norm on \mathcal{H}

$$\|\beta\|_{\mathcal{P}} = \inf \left\{ \sum_{m=1}^M \|v_m\| : v_m \in \mathcal{H}, \sum_{m=1}^M P_m v_m = \beta \right\}, \beta \in \mathcal{H},$$

whose dual norm is given by

$$\|x\|_{\mathcal{P},*} = \max_{m=1}^M \|P_m x\|.$$

These are the norms considered in (Maurer and Pontil, 2012) and include among others the group lasso, overlapping groups and multiple kernel learning. In the case of multiple kernel learning, for example, P_m is just the projection onto the m -th RKHS. We are interested in the Rademacher complexity of the function class $\mathcal{F} = \{x \in \mathcal{H} \mapsto \langle \beta, x \rangle : \|\beta\|_{\mathcal{P}} \leq 1\}$. Now

$$\begin{aligned} \mathbb{E} \sup_{\|\beta\|_{\mathcal{P}} \leq 1} \left\langle \beta, \sum_i \epsilon_i x_i \right\rangle &= \mathbb{E} \left\| \sum_i \epsilon_i x_i \right\|_{\mathcal{P},*} = \mathbb{E} \max_m \left\| \sum_i \epsilon_i P_m x_i \right\| \\ &= \mathbb{E} \max_m \sup_{\|\beta\|=1} \sum_i \epsilon_i \langle \beta, P_m x_i \rangle = \mathbb{E} \max_m \sup_{f \in \mathcal{F}_m} \sum_i f(x_i), \end{aligned}$$

where $\mathcal{F}_m = \{x \in \mathcal{H} \mapsto \langle \beta, P_m x \rangle : \|\beta\| \leq 1\}$, so Lemma 2 can be applied. Using (5) strong and weak parameters are

$$\begin{aligned} \mathfrak{S} &= \frac{2}{n} \max_m \mathbb{E} \left\| \sum_i \epsilon_i P_m x_i \right\| \leq 2 \max_m \sqrt{\frac{\text{tr}(\hat{C}(P_m \mathbf{x}))}{n}}, \\ \mathfrak{W} &= \sqrt{\max_m \sup_y \frac{1}{n} \sum_{i=1}^n \langle y, P_m x_i \rangle^2} = \max_m \sqrt{\lambda_{\max}(\hat{C}(P_m \mathbf{x}))}. \end{aligned}$$

Lemma 2 yields the overall bound

$$\mathcal{R}(\mathcal{F}, \mathbf{x}) \leq 2 \max_m \sqrt{\frac{\text{tr}(\hat{C}(P_m \mathbf{x}))}{n}} + 8 \max_m \sqrt{\frac{\lambda_{\max}(\hat{C}(P_m \mathbf{x})) \ln M}{n}}$$

which improves over the bounds in (Maurer and Pontil, 2012; Kakade et al., 2012; Cortes et al., 2010), whenever $\max_m \lambda_{\max}(\hat{C}(P_m \mathbf{x}))$ is appreciably smaller than $\max_m \text{tr}(\hat{C}(P_m \mathbf{x}))$.

3.2. Generalities on Multitask Dictionary Learning

We first consider multitask feature learning in general (Baxter, 2000). In subsequent sections we give exemplifying bounds for three specific regularizers.

With inputs in some space \mathcal{X} and intermediate feature representations in some feature space \mathcal{X}' let \mathcal{G} be a class of feature maps $g : \mathcal{X} \rightarrow \mathcal{X}'$ and let \mathcal{F} be a class of vector valued functions $\mathbf{f} : \mathcal{X}' \rightarrow \mathbb{R}^T$. We study the vector valued function class

$$\mathcal{F} \circ \mathcal{G} = \{x \mapsto (f_1(g(x)), \dots, f_T(g(x))) : g \in \mathcal{G}, \mathbf{f} \in \mathcal{F}\}.$$

Now let $x_{ti} \in \mathcal{X}$ be the i -th example available for the t -th task, $1 \leq i \leq n$ and $1 \leq t \leq T$. The multitask Rademacher average is now

$$\mathcal{R}(\mathcal{F} \circ \mathcal{G}, \mathbf{x}) = \frac{2}{nT} \mathbb{E} \sup_{\mathbf{f} \in \mathcal{F}} \sup_{g \in \mathcal{G}} \sum_{t=1}^T \sum_{i=1}^n \epsilon_{ti} f_t(g(x_{ti})),$$

where the ϵ_{ti} are nT independent Rademacher variables. The purpose of bounding these averages is to obtain uniform bounds in $\mathcal{F} \circ \mathcal{G}$ on the average multitask error $\frac{1}{T} \sum_t \mathbb{E} f_t(g(x_t))$, in terms of its empirical counterpart when x_t is sampled iid to x_{ti} (see e.g. Ando and Zhang, 2005). If \mathcal{F} is finite then the above expression evidently has the form required for application of Lemma 2, which gives

$$\mathcal{R}(\mathcal{F} \circ \mathcal{G}, \mathbf{x}) \leq \mathfrak{S} + 8\mathfrak{W} \sqrt{\frac{\ln(|\mathcal{F}|)}{nT}}$$

with strong and weak parameters

$$\mathfrak{S} = \frac{2}{nT} \max_{\mathbf{f} \in \mathcal{F}} \mathbb{E} \sup_{g \in \mathcal{G}} \sum_{t,i} \epsilon_{ti} f_t(g(x_{ti})) \text{ and } \mathfrak{W} = \sqrt{\max_{\mathbf{f} \in \mathcal{F}} \sup_{g \in \mathcal{G}} \frac{1}{nT} \sum_{t,i} f_t(g(x_{ti}))^2}.$$

In some cases the vector valued functions in \mathcal{F} consist of unconstrained T -tuples of real valued functions chosen from some class \mathcal{F}_0 independently for each task, so that $\mathcal{F} = (\mathcal{F}_0)^T$. In this case $\ln |\mathcal{F}| = T \ln |\mathcal{F}_0|$, so the above bound becomes

$$\mathcal{R}(\mathcal{F} \circ \mathcal{G}, \mathbf{x}) \leq \mathfrak{S} + 8\mathfrak{W} \sqrt{\frac{\ln |\mathcal{F}_0|}{n}}.$$

Typically $\mathfrak{S} \rightarrow 0$ in the multitask limit $T \rightarrow \infty$. This highlights the role of the weak parameter \mathfrak{W} . It controls what is left over of the generalization error for fixed n if T is large.

For a more concrete setting let $\mathcal{X} = \mathcal{H}$ be a Hilbert space and for some fixed $K \in \mathbb{N}$ let \mathcal{D} be the set of all dictionaries $D = (d_1, \dots, d_K) \in \mathcal{H}^K$ satisfying $\|d_k\| \leq 1$ for each k . The intermediate representation space will now be $\mathcal{X}' = \mathbb{R}^K$ and the admissible feature maps are

$$\mathcal{G} = \{x \in \mathcal{H} \mapsto (\langle d_1, x \rangle, \dots, \langle d_K, x \rangle) : (d_1, \dots, d_K) \in \mathcal{D}\}.$$

For a compact set of matrices $\mathcal{W} \subset \mathbb{R}^{T \times K}$ we define the class $\mathcal{F}(\mathcal{W})$ as

$$\mathcal{F}(\mathcal{W}) = \left\{ y \in \mathbb{R}^K \mapsto \left(\sum_k W_{1k} y_k, \dots, \sum_k W_{Tk} y_k \right) : W \in \mathcal{W} \right\}.$$

For fixed dictionary $D = (d_1, \dots, d_K)$ and fixed ϵ the expression $\sum_{t,i} \epsilon_{ti} \sum_k W_{tk} \langle d_k, x_{ti} \rangle$ is linear in W and therefore attains its maximum at an extreme point $W^* \in \text{ext}(\mathcal{W})$. Thus $\mathcal{R}(\mathcal{F}(\mathcal{W}) \circ \mathcal{G}, \mathbf{x}) = \mathcal{R}(\mathcal{F}(\text{ext}(\mathcal{W})) \circ \mathcal{G}, \mathbf{x})$. But the set of extreme points $\text{ext}(\mathcal{W})$ is often finite in which case our method can be applied. In the sequel we give two examples. Another possibility is that \mathcal{W} has a reasonable finite approximation, for which we will also give an example.

3.3. Dictionary Learning with the Sparsity Norm

For matrices $W \in \mathbb{R}^{T \times K}$ we define the sparsity norm¹

$$\|W\|_{\wedge} := \max_{t=1}^T \sum_k |W_{tk}|$$

and consider the class of matrices $\mathcal{W}_{\wedge} = \{W \in \mathbb{R}^{T \times K} : \|W\|_{\wedge} \leq 1\}$. Observe that $\mathcal{F}(\mathcal{W}_{\wedge}) = (\mathcal{F}_{\text{Lasso}})^T$, where $\mathcal{F}_{\text{Lasso}}$ is the class given by linear functionals on \mathbb{R}^K with ℓ_1 -norm bounded by 1. One checks that the set of extreme points is

$$\text{ext}(\mathcal{W}_{\wedge}) = \left\{ W : W_{tk} = \sigma_t \delta_{\phi_t, k}, \sigma_t \in \{-1, 1\}^T, \phi_t \in \{1, \dots, K\}^T \right\},$$

where δ is the Kronecker delta. In words: W is an extreme point iff for each t there is only one nonzero $W_{t\phi_t} \in \{-1, 1\}$, all the other W_{tk} being zero. Now $\text{ext}(\mathcal{W}_{\wedge})$ is finite with cardinality $|\text{ext}(\mathcal{W}_{\wedge})| = (2K)^T$, so our method is applicable to give bounds for the class $\mathcal{F}(\mathcal{W}_{\wedge}) \circ \mathcal{G}$. We bound the strong parameter as

$$\begin{aligned} \mathfrak{S} &= \frac{2}{nT} \max_{W \in \text{ext}(\mathcal{W}_{\wedge})} \mathbb{E} \sup_{D \in \mathcal{D}} \sum_{t,i} \epsilon_{ti} \sum_k W_{tk} \langle d_k, x_{ti} \rangle \\ &\leq \frac{2}{nT} \sup_{D \in \mathcal{D}} \left(\sum_k \|d_k\|^2 \right)^{1/2} \max_{W \in \text{ext}(\mathcal{W}_{\wedge})} \mathbb{E} \left(\sum_k \left\| \sum_{t,i} W_{tk} \epsilon_{ti} x_{ti} \right\|^2 \right)^{1/2} \\ &\leq \frac{2\sqrt{K}}{nT} \max_{W \in \text{ext}(\mathcal{W}_{\wedge})} \left(\sum_{t,i} \left(\sum_k W_{tk}^2 \right) \|x_{ti}\|^2 \right)^{1/2} \leq \frac{2}{nT} \sqrt{K \sum_{t,i} \|x_{ti}\|^2} = 2\sqrt{\frac{K \text{tr}(\hat{C}(\mathbf{x}))}{nT}}, \end{aligned}$$

where $\hat{C}(\mathbf{x})$ is the total covariance operator for all the data accross all tasks. Observe that we used no special properties of the extreme points, in fact we only used $\|W_t\|_2 \leq 1$ for all $W \in \mathcal{W}$. For the weak parameter we find

$$\begin{aligned} \mathfrak{W}^2 &= \max_{W \in \text{ext}(\mathcal{W}_{\wedge})} \sup_{D \in \mathcal{D}} \frac{1}{nT} \sum_{t,i} \left(\sum_k W_{tk} \langle d_k, x_{ti} \rangle \right)^2 = \max_{\phi \in \{1, \dots, K\}^T} \sup_{D \in \mathcal{D}} \frac{1}{nT} \sum_{t,i} \langle d_{\phi_t}, x_{ti} \rangle^2 \\ &\leq \frac{1}{T} \sum_t \sup_{\|d\| \leq 1} \frac{1}{n} \sum_i \langle d, x_{ti} \rangle^2 = \frac{1}{T} \sum_t \lambda_{\max}(\hat{C}(\mathbf{x}_t)), \end{aligned}$$

1. Some authors (see e.g. [Kakade et al., 2012](#)) would call this the $1/\infty$ -, others (see e.g. [Negahban and Wainwright, 2008](#)) the $\infty/1$ -norm, depending on preference for either computational or typographical order. To avoid confusion we use the wedge \wedge and refer to it as the “sparsity norm”.

where $\hat{C}(\mathbf{x}_t)$ is the covariance of the data of task t . The bound is

$$\mathcal{R}(\mathcal{F}(\mathcal{W}_\wedge) \circ \mathcal{G}, \mathbf{x}) \leq 2\sqrt{\frac{K \operatorname{tr}(\hat{C}(\mathbf{x}))}{nT}} + 8\sqrt{\frac{(1/T) \sum_t \lambda_{\max}(\hat{C}(\mathbf{x}_t)) \ln(2K)}{n}}.$$

This result has already been announced in (Maurer et al., 2013).

3.4. Dictionary Learning with the Sharing Norm

We reverse the order of summation and maximum in the definition of the previous norm to obtain the sharing norm

$$\|W\|_\vee = \sum_k \max_t |W_{tk}|.$$

This norm (under the name $1/\infty$ norm) has been applied to multitask learning by various authors (see e.g. Liu et al., 2009). Statistical guarantees in the form of oracle inequalities for multivariate regression have been given by Negahban and Wainwright (2008). None of these studies consider dictionary learning. To apply our method we first observe that the extreme points of the unit ball $\mathcal{W}_\vee = \{W : \|W\|_\vee \leq 1\}$ are now of the form $W_{tk} = v_t \delta_{k^*,k}$ for some $\mathbf{v} \in \{-1, 1\}^T$ and some $k^* \in \{1, \dots, K\}$. We have $|\operatorname{ext}(\mathcal{W}_\vee)| = 2^T K$.

For the strong parameter we find, using (5),

$$\begin{aligned} \mathfrak{S} &= \frac{2}{nT} \max_{W \in \operatorname{ext}(\mathcal{W}_\vee)} \mathbb{E} \sup_{D \in \mathcal{D}} \sum_{t,i} \epsilon_{ti} \sum_k W_{tk} \langle d_k, x_{ti} \rangle = \frac{2}{nT} \max_{\mathbf{v}, k^*} \mathbb{E} \sup_{D \in \mathcal{D}} \sum_{t,i} \epsilon_{ti} v_t \langle d_{k^*}, x_{ti} \rangle \\ &= \frac{2}{nT} \mathbb{E} \left\| \sum_{t,i} \epsilon_{ti} x_{ti} \right\| \leq \sqrt{\frac{\operatorname{tr}(\hat{C}(\mathbf{x}))}{nT}}. \end{aligned}$$

Here v_t disappears in the third identity. It is absorbed by the Rademacher variables, because the maximization is outside the expectation. By the same token the supremum over the dictionary becomes a supremum over a single vector v with $\|v\| \leq 1$ which leads to the norm. For the weak parameter we find

$$\mathfrak{W}^2 = \max_{W \in \operatorname{ext}(\mathcal{W}_\wedge)} \sup_{D \in \mathcal{D}} \frac{1}{nT} \sum_{t,i} \left(\sum_k W_{tk} \langle d_k, x_{ti} \rangle \right)^2 = \sup_{\|d\| \leq 1} \frac{1}{nT} \sum_{t,i} \langle d, x_{ti} \rangle^2 = \lambda_{\max}(\hat{C}(\mathbf{x})).$$

The overall bound is thus

$$\mathcal{R}(\mathcal{F}(\mathcal{W}_\wedge) \circ \mathcal{G}, \mathbf{x}) \leq 2\sqrt{\frac{\operatorname{tr}(\hat{C}(\mathbf{x}))}{nT}} + 8\sqrt{\frac{\lambda_{\max}(\hat{C}(\mathbf{x}))}{n} \left(\ln 2 + \frac{\ln K}{T} \right)}.$$

It depends only very weakly on the number K of dictionary atoms and only in the second term. Also observe that the weak parameter is never larger than in case of the sparsity norm $\|\cdot\|_\wedge$, because $\hat{C}(\mathbf{x}) = 1/T \sum_t \hat{C}(\mathbf{x}_t)$ and $\lambda_{\max}(\cdot)$ is convex on the cone of positive semidefinite operators.

A disadvantage of the sharing norm as a penalty is, that it makes strong assumptions on the relatedness of the tasks in question, and that it is sensitive to outlier tasks.

3.5. Subspace Learning

The final norm considered is

$$\|W\|_S = \max_t \left(\sum_k W_{tk}^2 \right)^{1/2},$$

which provides an opportunity to demonstrate our method when the norm on W is not polyhedral. We let \mathcal{W}_S be the unit ball in $\|\cdot\|_S$ and require the dictionary to be orthonormal. This is the class of multitask subspace learning (Ando and Zhang, 2005), where the effective weight vectors, $v_t = \sum_k W_{tk} d_k$, are constrained to all lie in a subspace of dimension K and to have norm bounded by one. We will derive a bound which compares well with bounds derived from the much more advanced methods of noncommutative Bernstein inequalities (Maurer and Pontil, 2013).

To apply our trick we first construct a finite approximation of \mathcal{W}_S with the help of covering numbers. Let $\eta > 0$. By (Cucker and Smale, 2001, Prop. 5) we can find a subset $\mathcal{W}_0 \subset \mathcal{W}_S$ such that $\forall W \in \mathcal{W}_S, \exists V \in \mathcal{W}_0$ such that $\|W - V\|_S \leq \eta$ and $|\mathcal{W}_0| \leq (4/\eta)^{KT}$. For every $V \in \mathcal{W}_0$ let $\mathcal{W}_V = \{W \in \mathcal{W}_S : \|W - V\|_S \leq \eta\}$, so that

$$\mathcal{W}_S = \bigcup_{V \in \mathcal{W}_0} \mathcal{W}_V.$$

We apply Lemma 2. By orthonormality of the dictionary the weak parameter is

$$\mathfrak{W}^2 = \max_{W \in \mathcal{W}_S} \sup_{D \in \mathcal{D}} \frac{1}{nT} \sum_{t,i} \left\langle \sum_k W_{tk} d_k, x_{ti} \right\rangle^2 = \max_{\|v\| \leq 1} \frac{1}{nT} \sum_{t,i} \langle v, x_{ti} \rangle^2 = \lambda_{\max} \left(\hat{C}(\mathbf{x}) \right).$$

The strong parameter can be bounded by two terms,

$$\begin{aligned} \mathfrak{S} &= \frac{2}{nT} \max_{V \in \mathcal{W}_0} \mathbb{E} \sup_{W \in \mathcal{W}_V} \sup_{D \in \mathcal{D}} \sum_{t,i} \epsilon_{ti} \sum_k W_{tk} \langle d_k, x_{ti} \rangle \\ &\leq \frac{2}{nT} \max_{V \in \mathcal{W}_0} \mathbb{E} \sup_{D \in \mathcal{D}} \sum_{t,i} \epsilon_{ti} \sum_k V_{tk} \langle d_k, x_{ti} \rangle + \frac{2}{nT} \mathbb{E} \sup_{\|W\|_S < \eta} \sup_{D \in \mathcal{D}} \sum_{t,i} \epsilon_{ti} \sum_k W_{tk} \langle d_k, x_{ti} \rangle. \end{aligned}$$

The first term is bounded by $2\sqrt{K \operatorname{tr}(\hat{C}(\mathbf{x})) / (nT)}$ exactly as in Section 3.3. For the second term we again use orthonormality of the dictionary

$$\begin{aligned} &\frac{2}{nT} \mathbb{E} \sup_{\|W\|_S < \eta} \sup_{D \in \mathcal{D}} \sum_t \left\langle \sum_k W_{tk} d_k, \sum_i \epsilon_{ti} x_{ti} \right\rangle \\ &= \frac{2}{nT} \mathbb{E} \sup_{D \in \mathcal{D}} \sum_t \sup_{\|w\| \leq \eta} \left\langle \sum_k w_k d_k, \sum_i \epsilon_{ti} x_{ti} \right\rangle = \frac{2\eta}{nT} \sum_t \mathbb{E} \left\| \sum_i \epsilon_{ti} x_{ti} \right\| \\ &\leq \frac{2\eta}{T} \sum_t \sqrt{\frac{\operatorname{tr}(\hat{C}(\mathbf{x}_t))}{n}} \leq 2\eta \sqrt{\frac{(1/T) \sum_t \operatorname{tr}(\hat{C}(\mathbf{x}_t))}{n}} = 2\eta \sqrt{\frac{\operatorname{tr}(\hat{C}(\mathbf{x}))}{n}}, \end{aligned}$$

where we used (5) and Jensen's inequality. Putting everything together and taking the infimum over η we get the bound

$$\begin{aligned} \mathcal{R}(\mathcal{F}(\mathcal{W}_S) \circ \mathcal{G}, \mathbf{x}) &\leq 2\sqrt{\frac{K \operatorname{tr}(\hat{C}(\mathbf{x}))}{nT}} + \\ &\quad + \inf_{\eta > 0} \left(2\eta\sqrt{\frac{\operatorname{tr}(\hat{C}(\mathbf{x}))}{n}} + 8\sqrt{\frac{K\lambda_{\max}(\hat{C}(\mathbf{x})) \ln(4/\eta)}{n}} \right). \end{aligned}$$

If $\mathcal{H} = \mathbb{R}^d$ we may for example set $\eta = \sqrt{K/d}$ to obtain

$$\mathcal{R}(\mathcal{F}(\mathcal{W}_S) \circ \mathcal{G}, \mathbf{x}) \leq 2\sqrt{\frac{K \operatorname{tr}(\hat{C}(\mathbf{x}))}{nT}} + 8\sqrt{\frac{K\lambda_{\max}(\hat{C}(\mathbf{x})) \ln(16d/K)}{n}}.$$

This can be compared to the bound derived from the results on trace norm regularization in (Maurer and Pontil, 2013). The present bound gives a faster approach to the limit as $T \rightarrow \infty$, but a larger limit value.

3.6. The limit $T \rightarrow \infty$ in High Dimensions

If X_1, \dots, X_n are sampled iid with $\|X_i\| \leq 1$ then (see e.g. Maurer and Pontil, 2013, Theorem 7)

$$\mathbb{E}\sqrt{\lambda_{\max}(\hat{C}(\mathbf{X}))} \leq \sqrt{\lambda_{\max}(C(X_1))} + 4\sqrt{\frac{\ln \min(\dim(\mathcal{H}), n) + 1}{n}}.$$

Here $C(X_1)$ is the true covariance $C(X_1) = \mathbb{E}\hat{C}(X_1)$. This allows to re-express our results in terms of expected Rademacher complexities, for which bounds as in Theorem 1 exist (Bartlett and Mendelson, 2002).

Now consider multitask dictionary learning as in the last two examples, with data sampled from the uniform distribution on the unit sphere \mathcal{S}^{d-1} in \mathbb{R}^d . Even if our multitask model is appropriate, to achieve empirical error η we will likely need to work with margins of order η/\sqrt{d} which incurs a Lipschitz constant of \sqrt{d}/η . On the other hand $C(X_1)$ has trace 1 and largest eigenvalue $1/d$. Thus, for fixed n , as $T \rightarrow \infty$,

$$\mathbb{E}\sqrt{\lambda_{\max}(\hat{C}(\mathbf{X}))} \rightarrow \sqrt{\frac{1}{d}}$$

which cancels the contribution of the Lipschitz constant. Applied to the bound in Section 3.4, the ambient dimension disappears completely in this limit. For subspace learning as in the previous section it appears only in the logarithm. This unveils a mechanism how multitask learning can potentially overcome the curse of high dimensionality.

Acknowledgments

This work was supported in part by EPSRC Grant EP/H027203/1 and Royal Society International Joint Project 2012/R2.

References

- R. K. Ando, T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.
- A, Argyriou, R. Foygel, N. Srebro. Sparse prediction with the k -support norm. *Advances in Neural Information Processing Systems 25*, pages 1466–1474, 2012.
- F. R. Bach, G.R.G. Lanckriet and M. I. Jordan. Multiple kernels learning, conic duality, and the SMO algorithm. Proc. 21st International Conference on Machine Learning, 2004.
- P. L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3:463–482, 2002.
- J. Baxter. A Model of Inductive Bias Learning, *Journal of Artificial Intelligence Research*, 12:149–198, 2000.
- S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. Proc. 16th Annual Conference on Computational Learning Theory, pages 567–580, 2003.
- S. Boucheron, G. Lugosi, P. Massart. Concentration Inequalities using the entropy method, *Annals of Probability*, 31(3):1145–1712, 2003
- S. Boucheron, G. Lugosi, P. Massart. *Concentration Inequalities*, Oxford University Press, 2013.
- C. Cortes, M. Mohri, A. Rostamizadeh. Generalization bounds for learning kernels. Proc. 27th International Conference on Machine Learning (ICML 2010), 2010.
- F. Cucker and S. Smale. On the mathematical foundations of learning, *Bulletin of the American Mathematical Society*, 39(1):1–49, 2001.
- L. Jacob, G. Obozinski, J. P. Vert. Group Lasso with overlap and graph Lasso. Proc. 26th International Conference on Machine Learning (ICML 2009), pages 433–440, 2009.
- S. M. Kakade, S. Shalev-Shwartz, A. Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13:1865–1890, 2012.
- V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.
- M. Ledoux, M. Talagrand. *Probability in Banach Spaces*, Springer, 1991.
- M. Ledoux. *The Concentration of Measure Phenomenon*, AMS Surveys and Monographs 89, 2001.
- H. Liu, M. Palatucci, and J. Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. Proc. 27th International Conference on Machine Learning, pages 82–89, 2009.
- A. Maurer. Concentration inequalities for functions of independent variables. *Random Structures and Algorithms*, 29:121–138, 2006.

- A. Maurer. Thermodynamics and concentration. *Bernoulli*, 18(2):434–454, 2012.
- A. Maurer, M. Pontil. Structured sparsity and generalization. *Journal of Machine Learning Research*, 13:671–690, 2012.
- A. Maurer and M. Pontil. Excess risk bounds for multitask learning with trace norm regularization. Proc. 26th Annual Conference on Learning Theory, pages 55–76, 2013.
- A. Maurer, M. Pontil, B. Romera-Paredes. Sparse coding for multitask and transfer learning. Proc. 30th International Conference on Machine Learning, pages 343–351, 2013.
- C. McDiarmid. Concentration. In *Probabilistic Methods of Algorithmic Discrete Mathematics*, pages 195–248, Springer, 1998.
- C. A. Micchelli, J. M. Morales, M. Pontil. Regularizers for structured sparsity. *Advances in Computational Mathematics*, 38(3):455–489, 2013.
- S. Negahban, M. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of ℓ_1/ℓ_∞ regularization. *Advances in Neural Information Processing Systems 21* pages 1161–1168, 2008.
- Y. Ying and C. Campbell. Generalization bounds for learning the kernel problem. Proc. 22nd Conference on Learning Theory, 2009.
- M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 68(1): 49–67, 2006.

4. Appendix

For the reader's convenience we provide a self-contained proof of Theorem 4 with all the required intermediate results. Most of the material of this appendix can also be found in the book by [Boucheron et al. \(2013\)](#).

4.1. Concentration Inequalities and Proof of Theorem 4

Let $(\Omega, \Sigma, \mu) = \prod_{i=1}^n (\Omega_i, \Sigma_i, \mu_i)$ be a product of probability spaces. For an event $\mathcal{E} \in \Sigma$ we write $\Pr(\mathcal{E}) = \mu(\mathcal{E})$. We denote a generic member of Ω by $\mathbf{x} = (x_1, \dots, x_n)$. For $\mathbf{x} \in \Omega$, $1 \leq k \leq n$ and $y \in \Omega$ we use $\mathbf{x}_{k \leftarrow y}$ to denote the object obtained from \mathbf{x} by replacing the k -th coordinate of \mathbf{x} with y . That is

$$\mathbf{x}_{k \leftarrow y} = (x_1, \dots, x_{k-1}, y, x_{k+1}, \dots, x_n).$$

For $g \in L_\infty[\mu]$ we write $\mathbb{E}g$ for $\int_\Omega g(\mathbf{x}) d\mu(\mathbf{x})$, and for $k \in \{1, \dots, n\}$ we introduce the functions $\mathbb{E}_k[g]$, $\inf_k[g]$ and $\sup_k[g]$ by

$$\begin{aligned} \mathbb{E}_k[g](\mathbf{x}) &= \int_{\Omega_k} g(\mathbf{x}_{k \leftarrow y}) d\mu_k(y), \\ \inf_k g &= \inf_{y \in \Omega} g(\mathbf{x}_{k \leftarrow y}) \quad \text{and} \quad \sup_k g = \sup_{y \in \Omega} g(\mathbf{x}_{k \leftarrow y}), \end{aligned}$$

where \inf and \sup on the r.h.s. are essential infima and suprema. The functions $\mathbb{E}_k[g]$, $\inf_k[g]$ and $\sup_k[g]$ are in $L_\infty[\mu]$ and do depend on \mathbf{x} but not on x_k . Note that $\mathbb{E}_k[g]$ corresponds to the expectation conditional to all variables except x_k . We use $\|\cdot\|_\infty$ to denote the norm in $L_\infty[\mu]$.

We will use and establish the following concentration inequalities:

Theorem 5 *Let $F \in L_\infty[\mu]$ and define functionals A and B by*

$$\begin{aligned} A^2(F) &= \left\| \sum_{k=1}^n \left(\sup_k F - \inf_k F \right)^2 \right\|_\infty \\ B^2(F) &= \left\| \sum_{k=1}^n \left(F - \inf_k F \right)^2 \right\|_\infty \end{aligned}$$

Then for any $s > 0$

$$\begin{aligned} \text{(i)} \quad \Pr\{F > \mathbb{E}F + s\} &\leq e^{-2s^2/A^2} \\ \text{(ii)} \quad \Pr\{F > \mathbb{E}F + s\} &\leq e^{-s^2/(2B^2)}. \end{aligned}$$

Part (i) is given in ([McDiarmid, 1998](#)). The inequality (ii) appears in different forms in various places [Boucheron et al. \(2003\)](#); [Ledoux \(2001\)](#). The constant 2 in the exponent appears first in ([Maurer, 2006](#)). We will use part (i) of the theorem to prove the following Gaussian concentration inequality, also known as the Tsirelson-Ibragimov-Sudakov inequality ([Ledoux and Talagrand, 1991](#); [Boucheron et al., 2013](#)).

Theorem 6 *Let $F : \mathbb{R}^n \rightarrow \mathbb{R}$ be Lipschitz with Lipschitz constant L and let $\mathbf{X} = (X_1, \dots, X_n)$ be a vector of independent random variables $X_i \sim \mathcal{N}(0, 1)$. Then for any $s > 0$*

$$\Pr \{F > \mathbb{E}F + s\} \leq e^{-s^2/(2L^2)}.$$

Before proving these results we show how they can be used to obtain Theorem 4.

Proof of Theorem 4. We first consider the bounded case and denote $F(\epsilon) = \sup_{\mathbf{z} \in A} \langle \epsilon, \mathbf{z} \rangle$. For any given ϵ let $\mathbf{z}(\epsilon) \in A$ denote a corresponding maximizer in the definition of F , so that $F(\epsilon) = \langle \epsilon, \mathbf{z}(\epsilon) \rangle$. Now fix a configuration ϵ . For any $j \in \{1, \dots, n\}$ and $\eta \in [-1, 1]$ recall that $\epsilon_{j \leftarrow \eta}$ denotes the configuration ϵ with ϵ_j replaced by η . Then for given j and η^* minimizing $F(\epsilon_{j \leftarrow \eta})$ we have

$$\begin{aligned} F(\epsilon) - \inf_{\eta \in [-1, 1]} F(\epsilon_{j \leftarrow \eta}) &= \langle \epsilon, \mathbf{z}(\epsilon) \rangle - \langle \epsilon_{j \leftarrow \eta^*}, \mathbf{z}(\epsilon_{j \leftarrow \eta^*}) \rangle \\ &\leq \langle \epsilon, \mathbf{z}(\epsilon) \rangle - \langle \epsilon_{j \leftarrow \eta^*}, \mathbf{z}(\epsilon) \rangle = (\epsilon_j - \eta^*) \mathbf{z}(\epsilon)_j \leq 2 \left| \mathbf{z}(\epsilon)_j \right|. \end{aligned}$$

It follows that

$$\sum_j \left(F(\epsilon) - \inf_{\eta \in [-1, 1]} F(\epsilon_{j \leftarrow \eta}) \right)^2 \leq 4 \|\mathbf{z}(\epsilon)\|^2 \leq 4 \sup_{\mathbf{z} \in A} \|\mathbf{z}\|^2$$

and the conclusion follows from Theorem 5 (ii).

For the normal case observe that the function $\mathbf{x} \in \mathbb{R}^n \mapsto \sup_{\mathbf{z} \in A} \langle \mathbf{x}, \mathbf{z} \rangle$ has Lipschitz constant $\sup_{\mathbf{z} \in A} \|\mathbf{z}\|$ and use Theorem 6. ■

4.2. A General Concentration Result

The proof of Theorems 5 and 6 is based on the entropy method (Ledoux, 2001; Boucheron et al., 2003, 2013). We first establish the following subadditivity property of entropy (Ledoux, 2001).

Theorem 7 *Suppose $g : \Omega \rightarrow \mathbb{R}$ is positive. Then*

$$\mathbb{E}[g \ln g] - \mathbb{E}[g] \ln \mathbb{E}[g] \leq \mathbb{E} \left[\sum_{k=1}^n (\mathbb{E}_k[g \ln g] - \mathbb{E}_k[g] \ln \mathbb{E}_k[g]) \right]. \quad (6)$$

To prove this we use the following lemma.

Lemma 8 *Let $h, g > 0$ be bounded measurable functions on Ω . Then for any expectation \mathbb{E}*

$$\mathbb{E}[h] \ln \frac{\mathbb{E}[h]}{\mathbb{E}[g]} \leq \mathbb{E} \left[h \ln \frac{h}{g} \right].$$

Proof Define an expectation functional \mathbb{E}_g by $\mathbb{E}_g[h] = \mathbb{E}[gh] / \mathbb{E}[g]$. The function $\Phi(t) = t \ln t$ is convex for positive t , since $\Phi'' = 1/t > 0$. Thus, by Jensen's inequality,

$$\mathbb{E}[h] \ln \frac{\mathbb{E}[h]}{\mathbb{E}[g]} = \mathbb{E}[g] \Phi \left(\mathbb{E}_g \left[\frac{h}{g} \right] \right) \leq \mathbb{E}[g] \mathbb{E}_g \left[\Phi \left(\frac{h}{g} \right) \right] = \mathbb{E} \left[h \ln \frac{h}{g} \right].$$

■

Proof of Theorem 7. Write $g/\mathbb{E}[g]$ as a telescopic product and use the previous lemma to get

$$\begin{aligned} \mathbb{E} \left[g \ln \frac{g}{\mathbb{E}[g]} \right] &= \mathbb{E} \left[g \ln \prod_{k=1}^n \frac{\mathbb{E}_1 \dots \mathbb{E}_{k-1}[g]}{\mathbb{E}_1 \dots \mathbb{E}_{k-1} \mathbb{E}_k[g]} \right] \\ &= \sum_k \mathbb{E} \left[\mathbb{E}_1 \dots \mathbb{E}_{k-1}[g] \ln \frac{\mathbb{E}_1 \dots \mathbb{E}_{k-1}[g]}{\mathbb{E}_1 \dots \mathbb{E}_{k-1}[\mathbb{E}_k[g]]} \right] \\ &\leq \sum_k \mathbb{E} \left[g \ln \frac{g}{\mathbb{E}_k[g]} \right] = \mathbb{E} \left[\sum_k \mathbb{E}_k \left[g \ln \frac{g}{\mathbb{E}_k[g]} \right] \right]. \end{aligned}$$

■

Fix some $F \in L_\infty[\mu]$. For any real β and $g \in L_\infty[\mu]$ define the thermal expectation $\mathbb{E}_{\beta F}[g]$ and for $1 \leq k \leq n$ the conditional thermal expectation $\mathbb{E}_{k,\beta F}[g]$ by

$$\mathbb{E}_{\beta F}[g] = \frac{\mathbb{E}[ge^{\beta F}]}{\mathbb{E}[e^{\beta F}]} \text{ and } \mathbb{E}_{k,\beta F}[g] = \frac{\mathbb{E}_k[ge^{\beta F}]}{\mathbb{E}_k[e^{\beta F}]}.$$

Also let $\sigma_{\beta F}^2[g]$ and $\sigma_{k,\beta F}^2[g]$ be the corresponding variances

$$\sigma_{\beta F}^2[g] = \mathbb{E}_{\beta F}[g^2] - (\mathbb{E}_{\beta F}[g])^2 \text{ and } \sigma_{k,\beta F}^2[g] = \mathbb{E}_{k,\beta F}[g^2] - (\mathbb{E}_{k,\beta F}[g])^2.$$

Note that $\mathbb{E}_{k,\beta F}[g]$ and $\sigma_{k,\beta F}^2[g]$ depend on \mathbf{x} but not on x_k . Also $\mathbb{E}_{k,\beta F}[g] = \mathbb{E}_{k,\beta(F+h)}[g]$ for any function h which does not depend on x_k . The Helmholtz free energy and its conditional counterpart are for $\beta \neq 0$

$$H(\beta) = \frac{1}{\beta} \ln \mathbb{E}[e^{\beta F}] \text{ and } H_k(\beta) = \frac{1}{\beta} \ln \mathbb{E}_k[e^{\beta F}].$$

Here we omit the dependence on F . Note that $\lim_{\beta \rightarrow 0} H(\beta) = \mathbb{E}[F]$.

Lemma 9 *We have*

$$H'(\beta) = \frac{1}{\beta^2} \int_0^\beta \int_t^\beta \sigma_{sF}^2[F] ds dt, \text{ and } H'_k(\beta) = \frac{1}{\beta^2} \int_0^\beta \int_t^\beta \sigma_{k,sF}^2[F] ds dt$$

Proof Define a function A by $A(\beta) = \ln \mathbb{E}[e^{\beta F}]$. Then $A(0) = 0$. It is easy to verify that $A'(\beta) = \mathbb{E}_{\beta F}[F]$ and $A''(\beta) = \sigma_{\beta F}^2[F]$. Thus

$$\mathbb{E}_{\beta F}[F] = A'(\beta) = A'(0) + \int_0^\beta A''(t) dt = \mathbb{E}[F] + \frac{1}{\beta} \int_0^\beta \int_0^\beta \sigma_{sF}^2[F] ds dt$$

and

$$\begin{aligned} \ln \mathbb{E}[e^{\beta F}] &= A(\beta) = \int_0^\beta A'(t) dt = \int_0^\beta \left(A'(0) + \int_0^t A''(s) ds \right) dt \\ &= \beta \mathbb{E}[F] + \int_0^\beta \int_0^t \sigma_{sF}^2[F] ds dt. \end{aligned}$$

Thus

$$H'(\beta) = \frac{1}{\beta} \mathbb{E}_{\beta F}[F] - \frac{1}{\beta^2} \ln \mathbb{E}[e^{\beta F}] = \frac{1}{\beta^2} \int_0^\beta \int_t^\beta \sigma_{sF}^2[F] ds dt,$$

which gives the first equation. The proof of the second is completely analogous. \blacksquare

We now give a general concentration result (see e.g. [Maurer, 2012](#)).

Theorem 10 *For any $\beta > 0$ we have the entropy bound*

$$H'(\beta) \leq \frac{1}{\beta^2} \mathbb{E}_{\beta F} \left[\sum_{k=1}^n \int_0^\beta \int_t^\beta \sigma_{k,sF}^2[F] ds dt \right] \quad (7)$$

and with $t > 0$ the concentration inequality

$$\Pr \{F - \mathbb{E}F > t\} \leq \exp \left(\beta \int_0^\beta H'(\gamma) d\gamma - \beta t \right). \quad (8)$$

Proof Substituting $g = e^{\beta F}$ in (6), dividing by $\beta^2 \mathbb{E}[e^{\beta F}]$ and using Lemma 9 we arrive at

$$\begin{aligned} H'(\beta) &\leq \mathbb{E}_{\beta F} \left[\sum_k \left(\frac{1}{\beta} \mathbb{E}_{k,\beta F}[F] - \frac{1}{\beta^2} \ln \mathbb{E}_k[e^{\beta F}] \right) \right] = \mathbb{E}_{\beta F} \left[\sum_k H'_k(\beta) \right] \\ &= \frac{1}{\beta^2} \mathbb{E}_{\beta F} \left[\sum_k \int_0^\beta \int_t^\beta \sigma_{k,sF}^2[F] ds dt \right], \end{aligned}$$

which is the first conclusion. Integrating H' from 0 to β , using $\lim_{\beta \rightarrow 0} H(\beta) = \mathbb{E}[F]$, and multiplying with β gives

$$\ln \mathbb{E}[e^{\beta F}] \leq \beta \mathbb{E}[F] + \beta \int_0^\beta H'(\gamma) d\gamma.$$

Subtract $\beta(\mathbb{E}[F] - t)$ and take the exponential to get

$$\mathbb{E}[e^{\beta(F - \mathbb{E}[F] - t)}] \leq \exp \left(\beta \int_0^\beta H'(\gamma) d\gamma - \beta t \right).$$

The second conclusion then follows from Markov's inequality. \blacksquare

4.3. Proofs of Theorems 5 and 6

With Theorem 10 at hand we can prove a number of concentration inequalities if we manage to bound the right hand side in (7). We then substitute in the second conclusion and optimize over β . At first the expression with the double integral and the thermal variances looks very cumbersome, but, as we shall see, it can often be bounded by comparatively simple methods. The bounded difference inequality, Theorem 5 (i) is obtained very easily, the proof of Theorem 5 (ii) is slightly more tricky.

Proof of Theorem 5 We prove (i). For fixed \mathbf{x} the thermal variance $\sigma_{k,sF}^2[F]$ is the variance of a function with values in the interval $[\inf_k F, \sup_k F]$, so that

$$\sigma_{k,sF}^2[F] \leq \frac{1}{4} \left(\sup_k F - \inf_k F \right)^2.$$

The double integral then just gives a factor of $\beta^2/2$. Now sum over k and bound the expectation $\mathbb{E}_{\beta F}$ by the $\|\cdot\|_\infty$ -norm to obtain

$$H'(\beta) \leq \frac{1}{8} \left\| \sum_k \left(\sup_k F - \inf_k F \right)^2 \right\|_\infty = \frac{A^2(F)}{8}.$$

(8) then gives

$$\Pr\{F - \mathbb{E}F > t\} \leq \exp\left(\frac{\beta^2}{8} A^2(F) - \beta t\right)$$

and substitution of $\beta = 4t/A^2(F)$ gives the result.

To prove part (ii) first note that for any expectation and any real function g we have $\sigma^2[g] = \min_{t \in \mathbb{R}} \mathbb{E}[(g - t)^2] \leq \mathbb{E}[(g - \inf g)^2]$. Applied to the conditional thermal variance this translates to

$$\sigma_{k,\beta F}^2[F] \leq \mathbb{E}_{k,\beta F} \left[\left(F - \inf_k F \right)^2 \right]. \quad (9)$$

We now claim that the right hand side above is a nondecreasing function of β . To see this write $h = F - \inf_k F$ and define a real function ξ by $\xi(t) = (\max\{t, 0\})^2$. Since $h \geq 0$ we have

$$\mathbb{E}_{k,\beta F} \left[\left(F - \inf_k F \right)^2 \right] = \mathbb{E}_{k,\beta(F - \inf_k F)} \left[\left(F - \inf_k F \right)^2 \right] = \mathbb{E}_{k,\beta h} [\xi(h)].$$

Here we used $\mathbb{E}_{k,\beta(F+h)} = \mathbb{E}_{k,\beta F}$ whenever g is independent of x_k . A straightforward computation shows

$$\frac{d}{d\beta} \mathbb{E}_{\beta h} [\xi(h)] = \mathbb{E}_{\beta h} [\xi(h) h] - \mathbb{E}_{\beta h} [\xi(h)] \mathbb{E}_{\beta h} [h] \geq 0,$$

where the last inequality uses the well known fact that for any expectation $\mathbb{E}[\xi(h) h] \geq \mathbb{E}[\xi(h)] \mathbb{E}[h]$ whenever ξ is a nondecreasing function. This establishes the claim.

Together with (9) this implies that for $0 \leq s \leq \beta$ we have

$$\sigma_{k,sF}^2[F] \leq \mathbb{E}_{k,sF} \left[\left(F - \inf_k F \right)^2 \right] \leq \mathbb{E}_{k,\beta F} \left[\left(F - \inf_k F \right)^2 \right],$$

so, using Theorem 10 again,

$$\begin{aligned} H'(\beta) &\leq \frac{1}{\beta^2} \mathbb{E}_{\beta F} \left[\sum_{k=1}^n \int_0^\beta \int_t^\beta \sigma_{k,sF}^2[F] ds dt \right] \leq \frac{1}{2} \mathbb{E}_{\beta F} \left[\sum_{k=1}^n \mathbb{E}_{k,\beta F} \left(F - \inf_k F \right)^2 \right] \\ &= \frac{1}{2} \mathbb{E}_{\beta F} \left[\sum_{k=1}^n \left(F - \inf_k F \right)^2 \right] \leq \frac{B^2(F)}{2}, \end{aligned}$$

where we used the identity $\mathbb{E}_{\beta F} \mathbb{E}_{k,\beta F} = \mathbb{E}_{\beta F}$. Then (8) gives

$$\Pr \{F - \mathbb{E}F > t\} \leq \exp \left(\frac{\beta^2 B^2(F)}{2} - \beta t \right)$$

and substitution of $\beta = t/B^2(F)$ gives the result. \blacksquare

Finally we use the bounded difference inequality, Theorem 5 (i), to prove the Gaussian Concentration inequality.

Proof of Theorem 6. By an easy approximation argument using convolution with Gaussian kernels of decreasing width it suffices to prove the result if the function F is in C^∞ with $|(\partial^2/x_i^2) F(\mathbf{x})| \leq B$ for all $\mathbf{x} \in \mathbb{R}^n$ and $i \in \{1, \dots, n\}$, where B is a finite, but potentially very large, constant. For $K \in \mathbb{N}$ let $X_i^{(K)}$ be the random variable

$$X_i^{(K)} = \frac{1}{\sqrt{K}} \sum_{k=1}^K \epsilon_{ik},$$

where the ϵ_{ik} are independent Rademacher variables, and define the random vector $\mathbf{X}^{(K)}$ accordingly. We write $G(\epsilon) = F(\mathbf{X}^{(K)})$ and set about to apply Theorem 5 (i) to the random variable $G(\epsilon)$ by bounding the variation in the epsilon components.

Fix a configuration ϵ with corresponding vector $\mathbf{X}^{(K)}$. For each $i \in \{1, \dots, n\}$ we introduce the real function $F_i(x) = F(\mathbf{X}_{i \leftarrow x}^{(K)})$. Since F is C^∞ we have for any $t \in \mathbb{R}$

$$F_i(x+t) - F_i(x) = tF_i'(x) + \frac{t^2}{2}F_i''(s)$$

for some $s \in \mathbb{R}$, and by the Lipschitz condition and the bound on $|F_i''|$

$$\begin{aligned} (F_i(x+t) - F_i(x))^2 &= t^2 (F_i'(x))^2 + t^3 F_i'(x) F_i''(s) + \frac{t^4}{4} (F_i''(s))^2 \\ &\leq t^2 (F_i'(x))^2 + |t|^3 LB + \frac{t^4}{4} B^2. \end{aligned}$$

Now fix a pair of indices (i, j) with $i \in \{1, \dots, n\}$ and $k \in \{1, \dots, K\}$. Since ϵ_{ik} can only have two values, one of which must be $X_i^{(K)}$, we must have

$$\begin{aligned} \left(\sup_y G(\epsilon_{(i,k) \leftarrow y}) - \inf_y G(\epsilon_{(i,k) \leftarrow y}) \right)^2 &= \left(F_i \left(X_i^{(K)} \pm \frac{2}{\sqrt{K}} \right) - F_i(X_i^{(K)}) \right)^2 \\ &\leq \frac{4(F_i'(X_i^{(K)}))^2}{K} + \frac{8LB}{K^{3/2}} + \frac{4B^2}{K^2}. \end{aligned}$$

Summing over k and i and then taking the supremum over ϵ we obtain

$$A(G)^2 \leq 4L^2 + \frac{8nLB}{K^{1/2}} + \frac{4nB^2}{K}.$$

From Theorem 5 (i) and $F(\mathbf{X}^{(K)}) = G(\epsilon)$ we conclude that

$$\Pr \left\{ F(\mathbf{X}^{(K)}) - \mathbb{E}F(\mathbf{X}^{(K)}) > s \right\} \leq \exp \left(\frac{-s^2}{2L^2 + 4nLB/K^{1/2} + 2nB^2/K} \right).$$

The conclusion now follows from the central limit theorem since $\mathbf{X}^{(K)} \rightarrow \mathbf{X}$ weakly as $K \rightarrow \infty$. ■